All Sciences Proceedings

1st International Conference on Recent and Innovative Results in Engineering and Technology

August 16-18, 2023 : Konya, Turkey

ICRIRET

All Sciences Proceedings
http://as-proceeding.com/

https://as-proceeding.com/index.php/icriret

Scraping of Vulnerable Physical Environment Images in Shodan Images in Turkey for Dataset

Fehmi Özkaner, Kıyas Kayaalp

Department of Computer Technologies Isparta University of Applied Sciences Isparta, Turkey

fehmi.ozkaner@gmail.com

Abstract – In this study, a scraping software has been developed with python in order to use the physical space images in Shodan images section as a dataset for CNN based artificial intelligence applications. With this software, the images listed in the Shodan images section within the scope of the query "country:en port:554" are collected by web scraping technique. The scraping software scraped and stored the images obtained as a result of the query every 3 days for an average of two months. As a result, 2,220 images were obtained to be used as a dataset in CNN-based classification applications.

Keywords – Shodan, Scraping, Vulnerable Physical Environment

I. INTRODUCTION

Shodan (Sentient-Hyper-Optimised-Data-Access-Network) is a search engine. However, unlike Google, Bing, and Yahoo, Shodan crawls the web for devices such as printers, security cameras, and routers connected to the internet. One of its main purposes is to identify security vulnerabilities. Shodan can connect to security cameras, traffic lights, baby monitors, and many other vulnerable devices. The IP address learned by pinging a site can be searched in Shodan to learn the ports, services, and general information about the site [1].

IP Camera is an imaging system that captures the signals received from sensors and converts these images into digital data and transmits them to authorized users with the help of a port. The protocol used to obtain real-time images from the IP Camera is RTSP (Real Time Streaming Protocol). RTSP allows you to obtain a live video stream from your camera and view the image from the network and different devices and programs. Its primary uses are to transmit video from a camera to an NVR, imaging software, and even home automation solutions. The RTSP port uses port 554 as standard.

II. RELATED WORK

Georgescu's work focused on enhancing the diagnostic and detection process for potential vulnerabilities in Internet of Things (IoT) systems through the utilization of a named entity recognition (NER) based solution. The proposed system was developed as a semantic indexing tool to catalog existing vulnerabilities, serving as an invaluable resource for security management experts. By incorporating this system, users can effortlessly identify potential vulnerabilities in IoT devices. The proposed solution combines ontologies and NER techniques to achieve a high level of automation, ensuring a self-maintaining and up-to-date system regarding vulnerabilities and common exposure information. The research involved identifying a total of 312 CVEs (common vulnerabilities and risks) specific to the IoT domain. The objective was to create an organized and structured framework tailored to user requirements by automatically monitoring and filtering IoT-related cybersecurity information. The paper's novelty lies in the adoption of a domain-oriented approach for IoT systems and the development of an object recognition algorithm dedicated to addressing cybersecurity concerns in the IoT domain. An automated data scraper

facilitates the selection of relevant information from the internet, streamlining the size of the database[2].

In their study, Walia et al. presented use cases and applications for cyber security in IoT-based scenarios, with effective cases and overall performance. There are smart devices and webcams that are connected to the IoT network environment with computing addresses. These information-based webcams are indexed by specialized search engines and Shodan is one of them. Shodan (shodan.io) is the major search engine that indexes and saves the data of computational webcams. By putting only webcams, security, and integrity are not implemented without ensuring the security of IT addresses. Shodan has retrieved data on vulnerable webcams and IoT devices. It was stated that it is necessary to integrate advanced strong passwords to ensure the security and privacy of webcams and IoT devices. When webcams are not given strong passwords, remote webcams can be accessed using the Python framework. As research academics and practitioners continue to work on research projects, Python-based tools advocate the ease of programming and high performance [3].

Arnaert et al. highlight the challenges faced by Information Technology (IT) and security administrators in managing the security of Internet-connected devices. These administrators often resort to search engines to identify device vulnerabilities. However, the authors argue that these search engines, while powerful, present difficulties in terms of complex query syntaxes and managing large quantities of results. This poses a time-consuming and challenging task for IT and security managers. In order to address these issues, the authors propose an ontology in their research to mitigate complexity and enhance the search engine results. The ontology aids administrators in detecting vulnerable devices effectively. Their primary objective is to streamline the results obtained from Censys and Shodan by reducing semantic complexity, improving the relevance of vulnerable objects, and enhancing the usability of these tools. The authors developed a Python program to implement their ontology, allowing them to successfully test its functionality with both Shodan and Censys. Through these tests, they obtained a concise list of public IP addresses associated with open-type protocols. The initial tests served as validation for their proposal, warranting further exploration [4].

Novianto et al. introduced an initiative known as Project SHINE, which aimed to gather comprehensive information on all Autonomous System Numbers (AS Numbers) in Indonesia using the Shodan search engine. The project focused on collecting data from Internet-connected devices, with a particular emphasis on identifying critical systems directly accessible on the Internet. The 2014 reports from Project SHINE revealed the identification of 211 countries based on the collected IP addresses. The top countries included the United States with 616,994 IP addresses (33.75%), Germany with 280,248 IP addresses (15.33%), China with 112,114 IP addresses (6.13%), Korea with 99,856 IP addresses (5.46%), and the United Kingdom with 66,234 IP addresses (3.62%). The authors successfully compiled a dataset within Shodan containing comprehensive information on AS Numbers from Indonesia. Within this dataset, they found 272,457 records of port information, 2,141 records of operating systems, 98,152 records of product services, 113,358 records of domains, 272,461 records of IP addresses, and 271,731 records of organization names, including Internet Service Providers, agencies, and universities. By utilizing a clustering algorithm, the authors identified four classes of AS Numbers based on their exposure level in Shodan. Class 0 indicated a lack of information about the AS Number in Shodan, while Class 1 denoted low information availability with nine AS Numbers falling into this category. The authors argue that these classes play a vital role in alerting organizations managing AS Numbers about the security of their systems [5].

In their study, Angelelli et al. proposed a quantitative regression model as a basis for ranking cyber vulnerability impact dimensions. In order to support statistical modeling for informed cyber risk assessment and threat intelligence, the prioritization of these impact dimensions by quantitative level is discussed. The sources supporting regulators and companies to make more

informed decisions for cyber risk assessment are heterogeneous and include severity ratings produced by organizations (notably the National Institute of Standards and Technology - NIST and Computer Security Incident Response Teams - CSIRTs) as well as reports, expert assessments and data from web resources and databases. The decision maker can choose different (official and unofficial) sources to prioritize both defensive and offensive actions depending on their objectives. In this contribution, we investigate the role of risk level attribution in ranking cyber vulnerabilities based on a set of observations from different databases. The study is a preliminary work on statistical modeling for threat intelligence, with special emphasis on sources of information on cyber vulnerabilities and risk acceptance/avoidance effects. On the other hand, deeper research is needed to explore the relationship between statistical (partial) ranking models, formal decision criteria, and sources of uncertainty that may lead to multiple prioritization in the cybersecurity domain [6].

Meknassi et al. presented in their work a methodology or automatic identification and collection of urban data that can assist in the initiation and implementation of smart city projects. The proposed method combines various techniques such as web scraping, search engines for linked objects such as Shodan, semantic analysis, and data fusion through APIs and Resource Description Framework (RDF). They aimed to help cities overcome the challenge of identifying, collecting, and combining data in smart cities due to their large, fragmented, multi-sourced, and heterogeneous nature. They proposed a comprehensive methodology that combines different technologies (WebScraping, search engine for IoT, APIs, AI, and semantic analysis) for the identification and collection of big urban data. For the semantic analysis of urban data, results are obtained in the form of a two-dimensional graph. This analysis allowed the identification of objects in the images. It successfully demonstrated some examples for the identification of the number of people and crowd density in the image [7].

In their work, Johnson et al. aimed to create a tool that would make it easier to use the vast amount of data on publicly available and discoverable devices, as well as the vast amount of information to be obtained on the security and vulnerability of control system devices. They created a tool to query databases such as Shodan. From the vulnerability information, a risk assessment was made to see if the device was possibly targeted or possibly compromised. Currently, many ICS devices that control the necessary infrastructure are not always properly protected. The study aimed to help ICS device operators better understand the cyber security risks associated with their devices. The detailed information in the study is intended to be used to help ICS operators make decisions about whether they need to increase the security of their devices [8].

Fernandes et al. conducted a study where they presented comprehensive guidelines for detecting vulnerabilities in IoT devices, specifically focusing on two prevalent cybersecurity concerns: the utilization of weak security mechanisms and the absence of appropriate security configurations. The researchers also elaborated on the process of automating vulnerability assessments for IoT devices using Shodan, a powerful search engine designed to explore the Internet and identify connected IoT devices. Furthermore, they put forth a methodology centered around the practical application of Shodan as a tool for teaching IoT cybersecurity, based on the analysis of real-world usage scenarios [9].

Within the scope of this study, the risks and consequences for companies, operators, and society due to the exposure of industrial facilities and Critical Infrastructures (CI) to failures or disruptions, power outages, or interruptions in production are analyzed based on the Shodan search engine. The consequences of IoT attacks on industrial facilities were found to be less predictable than attacks on traditional IoT systems, and increased network connectivity beyond the limits of systems that are no longer isolated, posing risks to IoT security and thus to the availability of these systems. A large number of SCADA systems and Industrial Control Systems are inadequately protected in terms of information technology, as the lack of authentication mechanisms greatly simplifies these attacks. Using

the Shodan search engine, it was found that these insufficiently secure control systems can be accessed over the Internet.

Bodenheim et al. conducted a study to assess the indexing functionality of Shodan, focusing on the crawl routine, crawl frequency, and timeliness of web database identification. The researchers recognized the widespread use of industrial control systems in critical infrastructure assets such as oil and gas pipelines, water distribution systems, power grids, nuclear power plants, and manufacturing facilities. To evaluate Shodan's indexing and querying capabilities, they deployed four Allen-Bradley ControlLogix PLCs in an Internet-facing configuration. Within a span of 19 days, Shodan successfully indexed and identified all four PLCs. The authors also discussed a potential mitigation strategy involving the manipulation of service banners to limit exposure to Shodan queries [10].

III. SHODAN

Developed in 2009 by John Matherly, a computer programmer, Shodan is a visual analysis tool for obtaining information by scanning the system in a wide area ranging from all objects connected to the internet and network devices such as various broadcast IP cameras systems, servers, web applications, routers and firewalls to the end units of industrial systems (SCADA).

Named after the artificial intelligence character in the video game "System Shock", Shodan can collect information from many central management and networked systems ranging from traffic light control and coordination systems to water facilities, power grids, and sewage management systems. It mostly collects this data through ports and end sensors connected to the internet. Shodan is a passive information collection tool. Shodan, which has many features; It is frequently used in open-source intelligence, penetration tests, or for information gathering by hackers.

Information about hundreds of millions of devices can be accessed from within the script/tool. Shodan has access to all the information collected with the API. All Shodan websites are fully supported by the public Shodan API, meaning that anything you can do through the website, you can do in code (Figure 1).

The Shodan Common line interface (CLI) allows you to automate the workflow or get the information you need efficiently without visiting the website.

Shodan has a series of commands for making enquiries. The commands are used to filter by the specified country code (country: tr), to filter by the city of the specified country when used with the country (city: Isparta), to filter by hostname or domain information (hostname: lawtech.com.tr), to search by the specified operating system (os:isletimsistemi), and to filter by port information (port:portnumber).

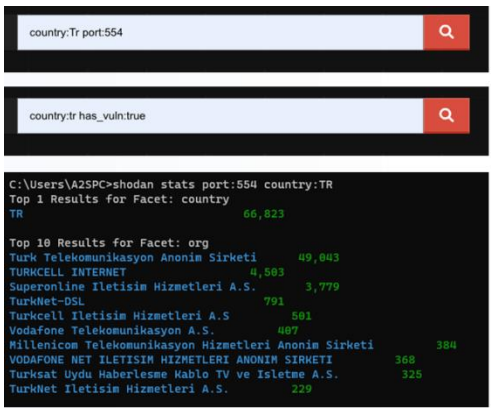

Figure 1. Shodan query example

Shodan collects screenshots for many different services, and some of its packages include access to a new search interface that makes browsing these screenshots much easier. In addition, Shodan performs OCR on these images, so you can also search for text within the images.

The search box in Shodan Images employs the identical syntax as the primary Shodan search engine. Its primary utility lies in filtering results based on an organization or network block. Additionally, the search box can be utilized to narrow down the types of images displayed. Shodan collects image data from five distinct sources, namely Remote Desktop (RDP), RTSP, Webcams, and X Windows. Each image source is associated with a different port or service, resulting in unique titles for each [1].

IV. SCRAPING SOFTWARE

A script has been developed with Python programming language to scrape the original images to be used for the dataset of artificial intelligence-based models from the SHODAN

search engine. SHODAN allows vulnerable and screenshotted IP addresses to pass through the API. The development stages of the web scraping software are given as follows.

Firstly, installation is done with the "pip install shodan" command to run shodan queries in the Python development environment. The API key must be obtained to use Shodan libraries in Python language. For this, API_key is obtained in the account section of the Shodan web page. In order to use the library and API_key in Python, the code line is added as follows.

```
from shodan import Shodan
api = Shodan('MY API KEY')
```

Shodan queries are made through the Shodan API. This query API method will be used for image scraping. The images taken from the 554th port of the security cameras in Turkey will be used for the dataset. For this process, we create an API query in Python language as follows.

```
Shodan search –fields  country:Tr port:554
has_screenshot:1 -screenshot.label:blank
```

With this query, images are displayed in the browser created using the from urllib.request import urlopen command in Python. For each image obtained, a URL in the form of www.shodan.io/host/81.215.71.142#554/image?p=554 is created in the Shodan search engine. In the last step, the images in the URL are batch scanned and downloaded every two days. The following code blocks are used in Python for this process.

```
[Python Code]
def download(url, pathname):
if not os.path.isdir(pathname):
os.makedirs(pathname)
response = requests.get(url, stream=True)
file_size = int(response.headers.get("Content-Length", 0))
filename = os.path.join(pathname, url.split("/")[-1])
progress = tqdm(response.iter_content(1024), f"İndiriliyor {filename}", total=file_size, unit="B", unit_scale=True, unit_divisor=1024)
with open(filename, "wb") as f:
for data in progress.iterable:
f.write(data)
progress.update(len(data))
```

The images listed in the Shodan images section within the scope of the query "country:en port:554" were collected by web scraping technique. Data were obtained for 2 months, provided that they were checked every three days. The total number of images obtained reached 2220 (Figure 2).

Figure 2. Sample images are taken from the Shodan

V. RESULT AND DISCUSSION

With the Python code written in the study, 2200 images were obtained from IP cameras with security vulnerabilities via Shodan. The images obtained through Shodan are continuously updated. The images on Shodan were checked to be in places where high security is required.

The aim of the study is to obtain a data set to be used in our next study, which is to identify the security vulnerabilities of places of strategic importance.

VI. CONCLUSION

In this study, a scraping software has been developed with Python in order to use the physical space images in the Shodan images section as a dataset from CNN-based artificial intelligence applications. With this software, the images listed in the Shodan images section within the scope of the query "country:en port:554" are collected by web scraping technique. Data were obtained for 2 months, provided that they were checked every

two days. The images were stored by scraping from the web page. As a result, 2220 images were obtained to be used as a dataset in CNN-based classification applications.

REFERENCES

[1] Matherly, J. (2015). Complete guide to shodan. Shodan, LLC.

[2] Georgescu, T. M., Iancu, B., & Zurini, M. (2019). Named-entity-recognition-based automated system for diagnosing cybersecurity situations in IoT networks. Sensors, 19(15), 3380.

[3] Walia, R., Oberoi, N., Kumar, A., & Singh, G. (2020). Digital Fingerprint and Security Aspects in Internet of Things Against Social Engineering Using Advanced Digital Forensics. Test Engineering and Management, 83(3), 4914-20.

[4] Arnaert, M., Bertrand, Y., & Boudaoud, K. (2016, July). Modeling vulnerable internet of things on shodan and censys: An ontology for cyber security. In Proceedings of the Tenth International Conference on Emerging Security Information, Systems and Technologies (SECUREWARE 2016) (pp. 299-302).

[5] Novianto, B., Suryanto, Y., & Ramli, K. (2021, March). Vulnerability analysis of internet devices from indonesia based on exposure data in shodan. In IOP Conference Series: Materials Science and Engineering (Vol. 1115, No. 1, p. 012045). IOP Publishing.

[6] Angelelli, M., & Catalano, C. A quantile regression ranking for cyber-risk assessment.

[7] Meknassi, H., & Shahrour, I. Intelligent data identification and collection for Smart City projects. MACHINE LEARNING & RISK ASSESSMENT IN GEOENGINEERING WROCŁAW, POLAND 25–27 OCTOBER 2021, 24.

[8] Johnson, G., Govindarasu, M., Ball, E., Foudree, G., Kammermeier, E., & Pals, R. Using Open Source Intelligence to Visualize Industrial Controller Risk

[9] Fernández-Caramés, T. M., & Fraga-Lamas, P. (2020). Teaching and learning iot cybersecurity and vulnerability assessment with shodan through practical use cases. Sensors, 20(11), 3048.

[10] Bodenheim, R., Butts, J., Dunlap, S., & Mullins, B. (2014). Evaluation of the ability of the Shodan search engine to identify Internet-facing industrial control devices. International Journal of Critical Infrastructure Protection, 7(2), 114-123.